به نام خدا

تمرین ۱، گزارش مربوط به دیتاست ۲

علیرضا نجاتی، ۹۸۲۲۲۱۰۴

مقدمه:

این دیتاست مربوط به اطلاعات آگهی های اجاره خونه های مختلف در مناطق مختلف کشور آلمان می باشد. در این دیتاست ویژگی های یک خونه و برخی از ویژگی های مربوط به آگهی آورده شده است. با استفاده از این داده می خواهیم قیمت اجاره یک خانه را با استفاده از دیگر داده ها پیش بینی و همچنین تاثیرگذاری ویژگی ها روی یکدیگر را بررسی کنیم.

توضیحات:

۸ سلول ابتدایی مربوط به اتصال به کگل، دریافت دیتاست، دسترسی کتابخانه های مورد نیاز و ساخت دیتافریم از روی دیتاست می باشد.

۱- تسک اول:

در تسک اول که پیش پردازش و پاکسازی داده ها در آن انجام شده است، ابتدا تعداد داده های null را برای هر ستون مشخص کرده و سپس آن ستون هایی که بیشتر از ۵۰ درصد داده های آن ها null هستند را حذف میکنیم (سلول های ۱۰ تا ۱۳).

سپس با توجه به اطلاعاتی که درون دیتاست قرار داشت، آن ستون هایی که اطلاعات مفیدی نداشتند و عملا قیمت اجاره خانه به آن ها ربطی ندارد را در سلول ۱۴ حذف کردیم.

در قسمت بعدی داده های عددی null را با میانگین ستون مربوطه پر می کنیم. از انجا که میخواهیم این قسمت را در ادامه به وسیله multiprocessing و dask انجام دهیم و runtime ها را باهم مقایسه کنیم، در این سلول runtime این عملیات محاسبه شده است که برابر با ۰,۰۵۵ است.

سپس به سراغ داده های null در ستون های categorical می رویم و آن ها را با داده با بیشترین تکرار در آن ستون جایگزین می کنیم (سلول ۱۸). دیگر داده پوچی در دیتاست وجود ندارد.

در ادامه داده هایی که احتمالا به اشتباه در دیتاست وارد شده اند را پیدا می کنیم. مثلا داده هایی که قیمت اجاره آن ها برابر با ۰ است و داده هایی که مساحت آن برابر با ۰ است را پیدا کرده و سطر مربوط به آن ها را حذف می کنیم (سلول های ۲۰ تا ۲۲).

در سلول ۲۴ داده های پرت عددی را به وسیله روش z_score پیدا می کنیم و آن ها را از دیتاست حذف می کنیم.
در سلول های ۲۶ و ۲۷ ستون های catrgorical که مقادیر مختلف زیادی را می گیرند را پیدا و آن ها را حذف می کنیم.
در سلول های ۲۸ و ۲۹ تعداد داده های دارای مشابهت را پیدا کرده سطر مربوط به آن ها را حذف می کنیم.

۲- تسک دوم:

در تسک دوم دریافت اطلاعات آماری و تحلیلی از دیتاست به همراه مصور سازی صورت گرفته است.
سلول ۳۰ اطلاعات آماری خوبی از هر ستون به ما می دهد.
در سلول ۳۱ و ۳۲ آمار و نمودار مربوط به منطقه ی خانه ها نمایش داده شده است که نشان می دهد منطقه Nordrhin_westfalen بیشترین رکورد ها را به خود اختصاص داده است.
در سلول ۳۳ و ۳۴ آمار و نمودار مربوط به شرایط خانه ها نمایش داده شده است که نشان می دهد well_kept بیشترین رکورد ها را به خود اختصاص داده است.
در سلول ۳۵ و ۳۶ آمار و نمودار مربوط به کیفیت داخلی خانه ها نمایش داده شده است که نشان می دهد کیفیت نرمال بیشترین رکورد ها را به خود اختصاص داده است.
در سلول ۳۷ و ۳۸ آمار و نمودار مربوط به نوع گرمایش خانه ها نمایش داده شده است که نشان می دهد که گرمایش مرکزی بیشترین رکورد ها را به خود اختصاص داده است.
در سلول ۴۱ برای ویژگی ها با تایپ Boolean نمودار توزیع رسم شده است.
نمودار مربوط به وابستگی ویژگی ها به یکدیگر در سلول ۴۲ رسم شده است(Correlation Plot). از این نمودار اینگونه برداشت می شود که قیمت اجاره بیشترین وابستگی را با هزینه خدمات، مساحت خانه و تعداد اتاق دارد. همچنین با توجه به اینکه totalRent با ویژگی های cellar، floor و garden وابستگی خیلی ناچیزی دارد، این ستون ها در سلول ۴۳ را از دیتاست حذف میکنیم.
نمودار پراکندگی مساحت و قیمت اجاره در سلول های ۴۴ و ۴۵ قرار دارد.
در سلول ۴۶ ارتباط بین منطقه و بازه قیمت پایه اجاره ترسیم شده است و همچنین نمودار مربوط به ارتباط نوع ساختمان و بازه پایه قیمت در سلول ۴۷ ترسیم شده است.

۳- تسک سوم:

تسک سوم مدلسازی قیمت ها براساس پارامتر های مختلف است. قبل از مدلسازی باید داده های categorical را به داده های عددی تبدیل کنیم برای این کار این متغیر ها را به متغیر های dummies تبدیل می کنیم (سلول ۴۹ تا ۵۲).

سپس ستون هدف را جدا کرده و بعد از آن داده های train و test را به نسبت ۸۰ به ۲۰ جدا میکنیم(سلول های ۵۴ تا ۵۶).

برای بهتر شدن نتیجه مدل در سلول ۵۷ داده ها را به روش MinMax اسکیل می کنیم.

سپس مدل Linear Regression را میسازیم و داده ها را به مدل می دهیم که مقدار خطای آن برابر با ۱۱۳,۲۸ و مقدار R^2_score آن که از فرمول RSS/TSS – ۱ بدست می آید برابر با ۰,۸۱ می شود.

و در سلول بعدی چند داده ی پیش بینی شده با داده اصلی مقایسه شده است (سلول های ۵۹ تا ۶۲)

۴- تسک چهارم:

تسک چهارم مربوط به استفاده از multiprocessing در پیش پردازش و پاکسازی داده هاست. در ابتدا گروه بندی داده ها بر حسب بازه قیمت پایه انجام می شود که در سلول ۶۵ این گروهبندی نمایش داده شده است. سپس قسمت پر کردن داده های null عددی را به وسیله multiprocessing در سلول ۶۶ انجام شده است که مقدار runtime آن برابر با ۱,۰۸۶ بدست آمد که از مقدار runtime در حالت عادی که در سلول ۱۴ بدست آمد بسیار بیشتر است.

۵- تسک پنجم:

تسک پنجم استفاده از dask برای مرحله پیش پردازش و پاکسازی داده هاست. در سلول های ۶۸ و۶۹ این کتابخانه نصب و import می شود. پس از آن dask dataFrame را از روی dataFrame اصلی میسازیم و سپس بخش پر کردن داده های null عددی را این بار با dask انجام می دهیم و runtime آن را اندازه می گیریم که برابر با ۰,۰۳۱ می شود و از حالت عادی کمتر است.